Rodríguez, J., Reguant, M., & Ortega, D. (2024). A practical case study of qualitative data analysis with IRaMuTeQ: lexicometric analysis of narratives of bisexual men and women. *Journal of Educational Research, 42*(2).
DOI: https://doi.org/10.6018/rie.560411

A practical case study of qualitative data analysis with IRaMuTeQ: lexicometric analysis of narratives of bisexual men and women

Un caso práctico de análisis de datos cualitativos con IRaMuTeQ: análisis lexicométrico de narrativas de hombres y mujeres bisexuales

Julio Rodríguez Rodríguez*[1] , Mercedes Reguant Álvarez* and Daniel Ortega Ortigoza**

*Department of Research Methods and Diagnosis in Education. University of Barcelona (Spain)
**Department of Applied Pedagogy. Universitat Autònoma de Barcelona (Spain)

Abstract

Qualitative data analysis is a complex matter. There are different forms of text analysis, including lexicometric analysis, which combines the richness of purely inductive analysis with the systematic rigour of quantitative analysis. Although there are controversies in the use of computer programmes for qualitative data analysis, this is not the case with lexicometry, due to its very nature. The aim of this methodological article is to present the potential of lexicometric analysis with IRaMuTeQ, using as an example the study of narratives of 80 bisexual men (15%) and women (85%) of different ages (M=26.55; SD=6.89). The results show the utility of IRaMuTeQ for the categorization of textual material, as it found six themes around which 77.50% of the narratives studied revolved. These include themes such as the absence of support, the difficulties for visibility, the bisexuality-homosexuality differentiation, the role of the environment, the fear of rejection, and the role of the LGTB community. In conclusion, IRaMuTeQ is presented as a useful piece of software for socio-educational research, allowing for the extraction of latent meanings in the textual material analysed using a structured and systematised working procedure such as the one described in this article.

Keywords: qualitative analysis; social education; lexicometry; methodology; research.

[1] **Correspondence:** Julio Rodríguez Rodríguez, julio.rodriguez.ro@ub.edu, Universitat de Barcelona, 08035 (Spain).

Resumen

El análisis de datos cualitativos representa una tarea compleja. Existen distintos tipos de análisis de textos, entre ellos, el análisis lexicométrico, el cual combina la riqueza del análisis puramente inductivo con el rigor de lo cuantitativo. Aunque existen ciertas controversias en la utilización de programas informáticos para el análisis cualitativo de datos, no es el caso de la lexicometría, debido a su propia naturaleza. El objetivo de este artículo metodológico es presentar las posibilidades del análisis lexicométrico con IRaMuTeQ, utilizando como ejemplo el estudio de las narrativas de 80 hombres (15%) y mujeres (85%) bisexuales (edad, M=26,55; DT=6,89). Los resultados muestran la utilidad de IRaMuTeQ en la clasificación del material textual, encontrando seis temáticas que aglutinan el 77,50% de las narrativas estudiadas, que tienen que ver con la ausencia de apoyos, las dificultades en la visibilidad, la diferenciación bisexualidad-homosexualidad, el papel del entorno, el miedo al rechazo, y el papel del colectivo LGTB. En conclusión, IRaMuTeQ se presenta como un software útil para la investigación socioeducativa, permitiendo la obtención de significados latentes en el material textual analizado, mediante un procedimiento de trabajo estructurado y sistematizado como el que se presenta.

Palabras clave: análisis cualitativo; educación social; investigación; lexicometría; metodología.

Introduction and objectives

Approaching human phenomena from the qualitative paradigm implies considering reality as a social construction, and highlighting language as an instrument of symbolisation (García Montejo, 2015). Language makes it possible to represent and communicate this shared world from the subjectivity of people, and it is qualitative research that is interested in the subjective perspective of people, and their interpretation and construction of reality (Ander-Egg, 2011; Sandín, 2003; Taylor and Bogdan, 1987). The use of qualitative methodology is "one of the most significant trends in educational research at the beginning of the 21st century" (Sabariego, 2004, p. 63).

Qualitative research is characterised by: being an iterative and emergent process, used primarily to explore, understand and produce new knowledge; containing a wealth of information about the phenomenon under study; seeking to produce knowledge to explain; and taking into account the context to understand what is being researched (Chandra and Shang, 2019).

Once the research design has been chosen, it is necessary to decide on the data collection strategies and the type of data analysis to be carried out. These decisions are closely related to the research paradigm, the nature of the problem and the objectives. In agreement with Trigueros et al. (2018), the variety of qualitative approaches leads to a lack of clear models in the analysis process. Therefore, they propose an integrative perspective when approaching data analysis, taking into account both the point of view and the interpretation of the narratives and implicit theories of the participants.

Qualitative data collection strategies include interviews, focus groups, participant observation, written documents (field diaries, summaries, WhatsApp or Twitter messages) and audiovisual records (audio, photography and video). Once this information is obtained, it is analysed and interpreted. This involves reducing that

information to meaningful units to be analysed. Qualitative analysis is an interpretative activity of the subjective processes, personal experiences, beliefs, etc. of the participants (Chernobilsky, 2009; Massot et al., 2004). Ander-Egg (2011) points out that "The purpose of analysis is to summarise the observations carried out in a way that provides answers to the research questions. The purpose of interpretation is to seek the broader meaning of the answers by working with other available knowledge." (p. 160).

There are different ways of approaching the analysis of qualitative data, depending on the research questions, the origin of these data, the methodological perspectives, and the methods of analysis themselves (Díaz Herrera, 2018; González Teruel, 2015; Miles and Huberman, 1994; Taylor and Bogdan, 1987): keyword counting, semantic network analysis, content analysis (quantitative and qualitative), analysis of grammatical structures, or conversations and narratives (Ryan and Bernard, 2000), grounded theory (Glaser and Strauss, 1967) and critical discourse analysis (van Dijk, 1999), or thematic analysis (Braun and Clarke, 2006). The analysis of qualitative data shares the aim of analysing and understanding in depth, taking into account, or not, the context in which qualitative data is generated. They seek, in short, to approach the latent meaning of narratives from the manifest content. As Abela (2002) points out, "any content of a text or an image can be interpreted in a direct and manifest way or in a hidden way of its latent meaning" (p. 2). In any case, the analysis of qualitative data is a complex task, which can complicate the approach to this task for both university students and novice researchers. A qualitative analysis requires a systematic and careful procedure, as it is possible to get lost in the wide world of qualitative data (Ary et al., 2010; Gibbs, 2012; Kalpokaite and Radivojevic, 2019). The volume of data in qualitative research exceeds that of quantitative research (Gibbs, 2012; Taylor and Bogdan, 1987). Tools are needed to facilitate data management and decision-making, and appropriate *software* can be a necessary aid (Huber and Marcelo, 1990; Richards and Richards, 2002; Saldaña, 2013).

In the diversity of existing types of textual data analysis, the statistical analysis of texts brings together "those procedures which, by calculating the occurrences of one or several basic verbal units, allow some kind of calculation to be made on the basis of the results obtained" (Jordà et al., 2000, p. 610). This type of analysis, also known as lexicometry, "groups together a series of methods that make it possible to reorganise the units present in a textual sequence and operate certain statistical analyses based on the quantification of the vocabulary resulting from a segmentation operated on the text" (Gil et al., 1994, p. 511). Lexicometry takes advantage of the possibilities of statistical analysis and computer science, combining the economy of processing with the precision granted to quantitative methods.

In this type of analysis "people elaborate representations that are configured in conceptions about different phenomena of the world. The description of the categories through the lexicometric method would show the most stable aspects of these ways of thinking" (Baccalá and de la Cruz, 2000, p.2). Basically, it allows us to identify "the "nuclear" ideas inferred from the characteristic words and their relations, that is, those inferred from the set of words associated in each category" (Baccalá and de la Cruz, 2000, p.2). This analysis is based on the frequency of different words and sequences of terms and their proximity to each other.

The statistical approach to textual data allows the latent meaning of the discourse to be obtained, and provides a complete, quantitative, and structured description of the text (Bécue et al., 1992; Peralta et al., 2020). Lexicometric analysis can: identify trends or

preferences of emerging terms in the thinking of an individual, as well as the collective thinking of a group; and make inferences based on different variables inherent to the participant(s) (Romero-Perez et al., 2018).

The opportunities offered by technology have also been present in the different types of qualitative data analysis. The acronym CAQDAS (*Computer-assisted Qualitative Data Analysis Software*) or QDA *software* (*Qualitative Data Analysis*), groups together computer programmes for the analysis of qualitative data (Carvajal, 2002; Farias and Montero, 2005). Trigueros et al. (2018) point out the antagonistic positions of supporters and detractors of the use of software tools for qualitative data analysis. As they skilfully point out, this is a sterile debate, since we should not lose sight of the fact that their use does not guarantee the quality of the analysis, although the goodness of their use depends on the ideological, epistemological and ethical stance of the researchers. In their opinion, the advantages of their use have to do with saving time, efficiency in information management, handling large amounts of data, and improving the quality of research, as well as teamwork and the management of the research process. However, these same authors point out a series of disadvantages: the transfer of quantitative design to qualitative research, the loss of the globality of the process, and the seduction by the quantification of data. Ultimately, this type of *software* facilitates analysis, although it does not replace the creativity and involvement of the researcher, nor the uncritical use of the software (Chandra and Shang, 2019; Gibbs, 2012; Hart and Achterman, 2017; Hernández-Sampieri and Mendoza, 2018). Therefore, the *software* is a technical aid, a means, but it is not the end of the research, and we believe that it also requires the active involvement of the researcher in all phases of the analysis process, as can be seen in the study presented in this article, in which we exemplify lexicometric analysis using the IRaMuTeQ *software*.

Commercial software includes Atlas-ti, Ethnograph or NVivo, although the price of the licences exceeds the economic possibilities of students and novice teams. There are other options based on so-called free (or open source) *software*, including AQUAD, or Weft QDA, to name a few, which are used in academic research (Ávalos et al., 2018; Gonçalves et al., 2021; Marina and Feliz, 2018).

There are also options for lexicometric analysis, among which are: CISIA's SPAD-T, from the Centre International de Statistique et d'Informatique Appliquées (France); LEXICLOUD, developed by the Laboratoire de Lexicometrie et Textes Politiques of the Ecole Normale Supérieure de Fontenay-St. Cloud (France); FRECON, developed by the Universitat de les Illes Balears (Spain); Orange, a data mining suite for lexicometric data mining. Cloud (France); FRECON, developed by the Universitat de les Illes Balears (Spain); Orange, a suite for data mining, developed by the Bioinformatics Laboratory, University of Ljubljana (Slovenia); and IRaMuTeQ, developed by the LERASS research laboratory (France).

IRaMuTeQ (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*) is a free and open source *software* for multidimensional text analysis, developed by Pierre Ratinaud at the University of Toulouse (Moreno and Ratinaud, 2015). IRaMuTeQ performs most analyses with minimal intervention by the researcher.

The aim of this methodological article was to present the possibilities of lexicometric analysis using the free *software* IRaMuTeQ, using as an example the study of the narratives of bisexual men and women.

Method: Qualitative analysis using the free software IRaMuTeQ.

The following is a procedure for using IRaMuTeQ that is useful for handling a large volume of qualitative data, and provides a different approach to traditional content analysis. This guidance may be useful for novice researchers who want to rely on software for qualitative data analysis. Camargo and Justo (2013) consider that IRaMuTeQ has much to contribute to qualitative analysis in studies linked to the Humanities or Social Sciences. These basic fundamentals do not replace the in-depth study of works such as Loubère and Ratinaud (2014), Moreno and Ratinaud (2015), Molina-Neira (2017), or Ruiz (2017) to use IRaMuTeQ to its full potential.

Before starting, it should be noted that IRaMuTeQ has some technical peculiarities: the free and open source *software* R and some additional libraries must be installed beforehand (Ruiz, 2017). Otherwise, IRaMuTeQ will not work (see any manual for a more detailed description). In this article, R 4.0.0, IRaMuTeQ version 0.7 alpha 2, and Windows 11 Pro have been used.

IRaMuTeQ offers various options for the analysis of qualitative data, and although the Top-Down Hierarchical Classification according to Reinert's method (Ratinaud, 2018; Sbalchiero, 2018) is one of the most common, other analysis possibilities should be explored (Souza et al., 2018). In this sense, IRaMuTeQ allows from a basic lexicographic analysis, the Top-Down Hierarchical Classification, the factorial analysis of correspondences, or the analysis of similarities. With all these analyses, the symbolic content of the data provided by the participants can be studied (Morales, 2021).

Preparing the data: basic assumptions

Huber and Marcelo (1990) pointed out that, in qualitative research, he provides an "overwhelming amount of rich and suggestive data" (p. 69), so qualitative material has to be prepared for further analysis: transcribing interviews, or preparing a table with answers to open-ended questions. Word processors and spreadsheets can be used for a fee, or free *software* such as OpenOffice can be used. Unlike other programmes, IRaMuTeQ does not allow you to work with different text files at the same time, so you will have the transcribed textual material in a single file.

Before starting the analysis process, it is key to check the text file thoroughly because IRaMuTeQ does not allow editing afterwards. The format, spelling and other formal aspects of the text document must be checked. Several authors (Camargo and Justo, 2016; Gourlay, 2019; Molina-Neira, 2017) suggest:

- Check typos and transcription errors.
- Delete the researcher's textual material (questions, annotations...) or keep them, marking them with an underscore (e.g. "_QuestionForDifficulties_").
- Avoid right-justified text, and do not use bold, italics or underlining.
- If acronyms are used, such as LGTB or LGTBIQ: always the same.
- Pay attention to synonymous words when the aim is not to address the phenomenon from a linguistic point of view. It may be of interest to keep words such as "homosexual" and "gay" separate.
- Do not use special characters in the text, such as " " & * @ # # € $ ... ( ) [ ] : ; -

- Compound words, such as "out of the wardrobe", should be joined with an underscore ("out_of_the_closet").
- All texts must be written in the same language.
- The numbers are kept in their numerical format.

Another particularity of the text document is that it must be saved in *.txt format, using the standard UTF-8 and LF encoding (Camargo and Justo, 2016). This can be done from OpenOffice.

Finally, take into account the placement and identification of textual material. The arrangement and identification of the material could follow different options (Camargo and Justo, 2016; Loubère and Ratinaud, 2014). If we have texts from semi-structured interviews[2] , one would be to place the content of the interviews as it is, as a single paragraph. In this case, in the first line we identify the participant, or the interview number, and the variables or characteristics we want to assign (age, gender, etc.). The format is: an asterisk followed by the variable name (e.g. *sex) and an underscore following for the modality of that variable (e.g. *sex_male).

**** *subject_078 *age_between25and35 *sex_h *gender_masc *marital_stay_unmarried

> Well, I've never felt like I've been in a wardrobe, my process has been more of openness than of discovery, so, just as I didn't have to sit my family down and tell them I was heterosexual, I found no reason to do the same with my bisexuality. So it's just come naturally and spontaneously. What I do have more respect for are friends in the village or other circles where I might be more exposed to comments. Luckily I have a family that wants us to be happy doing what we feel and with whom we feel, they are only there to support us with our projects and happiness.

Another option would be to identify the different questions or thematic blocks of each participant's interview. In this case, the procedure would be similar to the previous one, adding an additional line for each question or topic, with the format: dash, asterisk, topic, underscore and modality of that topic or theme (e.g. -*difficulties_topic).

**** *subject_225 *younger_age25 *sex_m *fem_gender * *unmarried_marital_state
-*difficulties_topic

[2] The data provided to exemplify this methodological article, with a sample of bisexual men and women, come from a broader study about homophobia in LGTBIQ+ people. The data presented here correspond to a part of the qualitative study that explores, through a survey of open questions, the difficulties and support persons in the process of becoming visible as an LGTBIQ+ person in this sample of bisexual men and women (*n=80*, 15% men and 85% women; *M=26*.55; *SD=6*.89) (Spain). The survey was distributed online among different LGTBIQ+ organisations in Catalonia, through the collaboration of key people from the different organisations. The purpose of the study was informed in writing, and the anonymity and confidentiality of all those who participated was guaranteed, with no data being collected that would allow them to be identified. The requirements for participation were: to be an LGTBIQ+ person, of legal age, and resident in Catalonia for at least the last year.

> Realising I wasn't straight, coming out and finding partners who didn't mind. People think that being bisexual means you'll cheat on them with someone of the opposite gender.
>
> -*support_topic
>
> My biggest support is my friends, who are also part of the collective, almost all of them.

It is important to pay attention to the format of the name assigned to the variables, as the program differentiates between upper and lower case letters: the variable *sex_Man would not be the same as the variable *sex_man. Once the text has been prepared and saved in *.txt format UTF-8 encoding, the lexicometric analysis can be started with IRaMuTeQ.

First steps: basic lexicographical analysis

Load the text file by clicking on the "text corpus" icon. In the screen that appears (figure 1), select the *utf8-all languages* option in the "Encoding" tab, and the language of the text in the "Language" tab. By default, the "Method of construction of the segments" is set to the occurrence method (it should be changed to *paragraphs* when we have short answers to open questions of a questionnaire). This action generates a first basic information about the textual corpus we have loaded (figure I). In this case, IRaMuTeQ indicates that this text corpus consists of 80 texts (corresponding to the 80 participants) and a total of 160 text segments. The occurrences are 2421 and the forms (generic words) 686. The words that appear with a frequency equal to 1, called hapax, are 437 in this corpus, representing 63.70% of the forms and 18.05% of the occurrences.

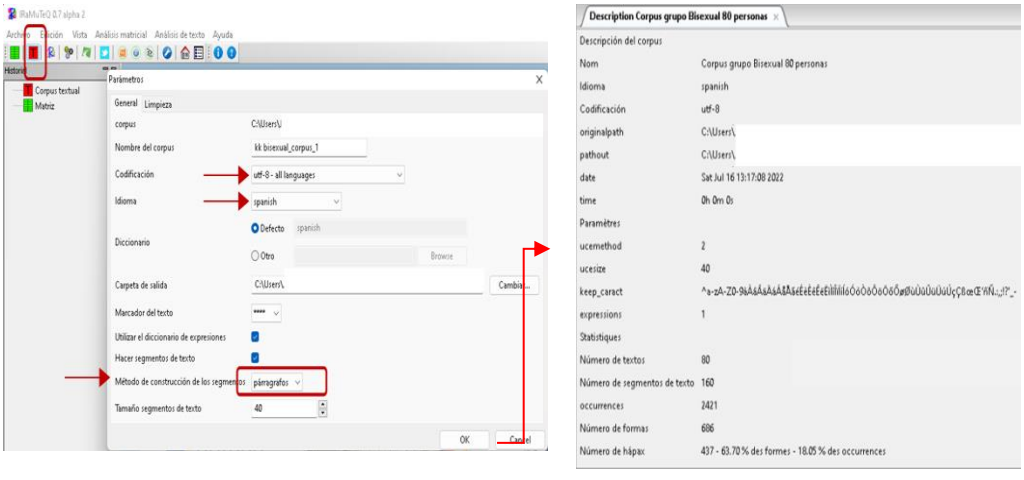

Figure 1. Initial parameters and descriptive statistics of the textual corpus.

As we pointed out, IRaMuTeQ allows different analyses of varying degrees of complexity. A first analysis is the lexicographic or textometric analysis which shows the frequencies of the words, the average frequency, and the active and supplementary forms. This first analysis is executed by clicking on the "Statistics" icon. The option "properties" (figure 2), allows us to select the type of words we want to include in the analysis.

For textual corpora linked to Psychology or Social Sciences, Camargo and Justo (2016) recommend activating only the boxes Adjectives (1), Uncommon Forms (1), Common Noun (1), Supplementary Noun (2), Verb (1) and Supplementary Verb (2), eliminating the "instrument words" (figure II). This action shows results with important information for further, more complex analyses. The tab "Active forms" (figure III) allows you to see the total list of forms (words), their frequency and the type of word, sorted according to their frequency. In our example (Figure IV), the selected word types are limited to common nouns, adjectives and verbs, as recommended by Camargo and Justo (2013).

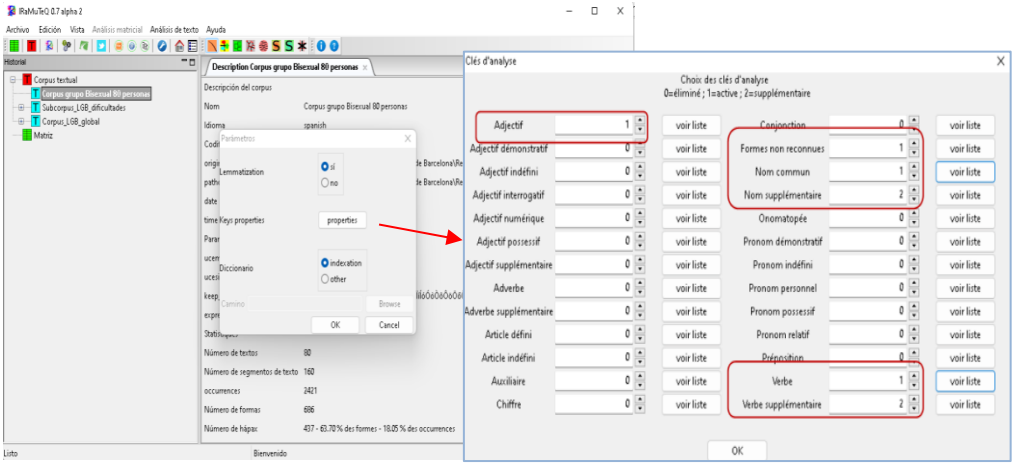

Figure 2. Selection of parameters and choice of analysis keys.

If we right-click on any of the words listed, a submenu allows us to see the "associated forms" and their frequency, and the "concordances" or text fragments in which these and the associated forms appear (figure 3). Associated forms are words of the same family. For example, the form "ir" has as associated forms the words "ido", "voy", "iba", "irme", "van", "vas" and "ir".

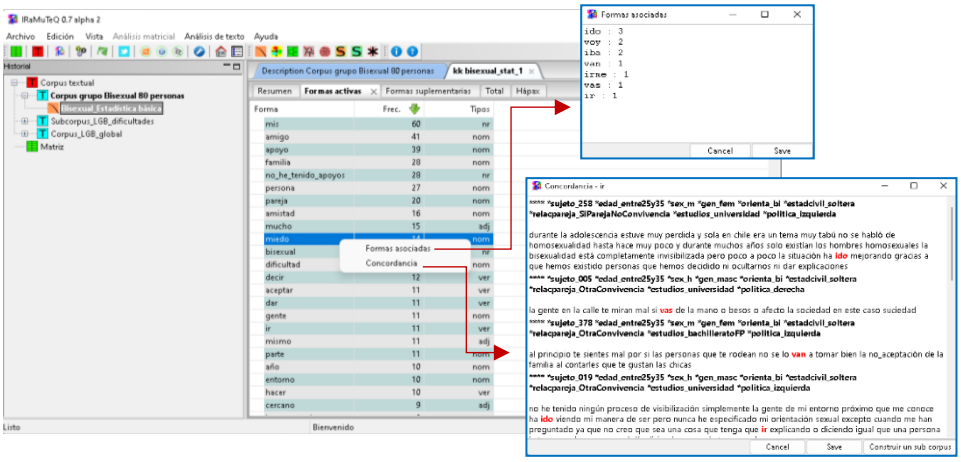

Figure 3. Summary, key shapes, complementary, total and hapax.

Another analysis is the Word Cloud, a graphical presentation (figure 4) of the words according to their frequency of occurrence in the text corpus, with a choice of active, supplementary or active and supplementary forms. The colour of the word cloud and of the background, the maximum and minimum size of the text, as well as the specific words that we want to appear within the active, supplementary, active and supplementary forms can be selected.

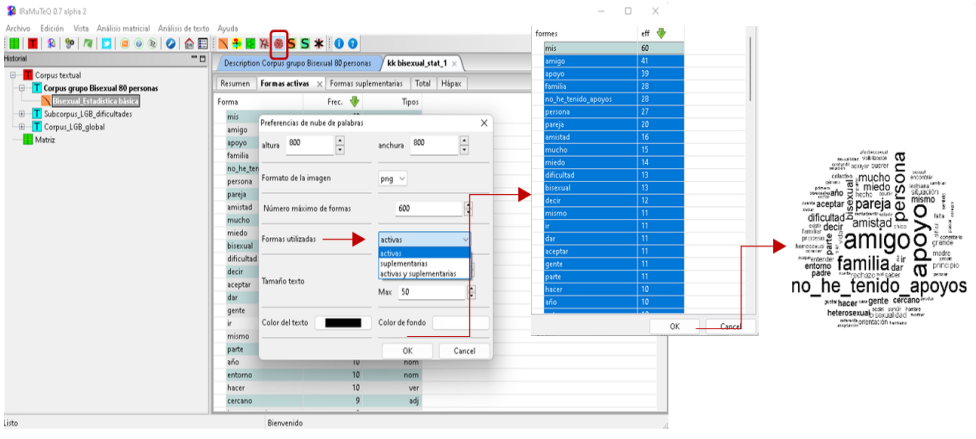

Figure 4. Word cloud generation process.

Advanced steps (I): Correspondence Factor Analysis (CFA)

Recall that the text file is organised in blocks consisting of at least one variable. Usually there is more than one variable (e.g. one identifying the participants, one related to age, one related to gender, etc.). Thus, a first variable is in our example the participant person, followed by age, sex, gender or marital status. To carry out the correspondence factor analysis (CFA), a variable is chosen that has at least three modalities (Camargo and Justo, 2016; Ruiz, 2018), such as marital status, divided in our example into: "single" (*single_marital_status), "separated/divorced" (*separated_marital_status) and "married" (*married_marital_status). As illustrated in figure 5, the form "have_not_had_supports" is more frequently found among single people (25) than among separated (1) or married (2) people. This same result can be represented graphically. This frequency histogram can also be made by groups of words.

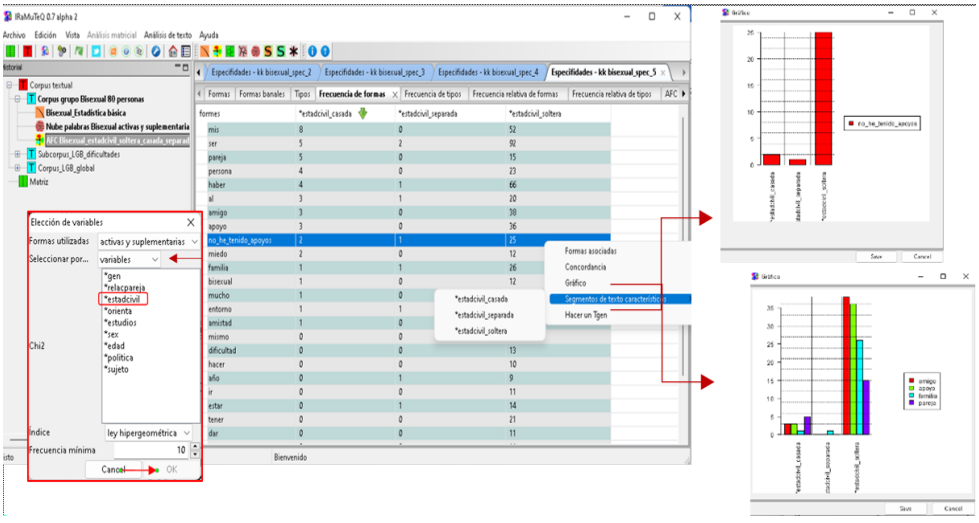

Figure 4. Correspondence Factor Analysis.

Advanced Steps (II): the Hierarchical Top-Down Classification (HCD)

Hierarchical Top-Down Classification (HCD) represents the relationship between reduced forms and text segments. That is, IRaMuTeQ groups the text fragments into classes or clusters of meanings, so that it classifies the forms into clusters of text that are similar in meaning. Each of the clusters has a number of characteristic words, ordered by their frequency of occurrence. This allows the typical vocabulary of each cluster to be contextualised.

It is important, on the one hand, to select the parameters of this analysis well (Figure 5), otherwise technical errors will occur. Thus, with short answer fragments, as in a questionnaire of open-ended questions, the option "Simple in the text" will be selected

(previously "paragraphs" will have been selected as the method of construction of the text segments) (Camargo and Justo, 2016).

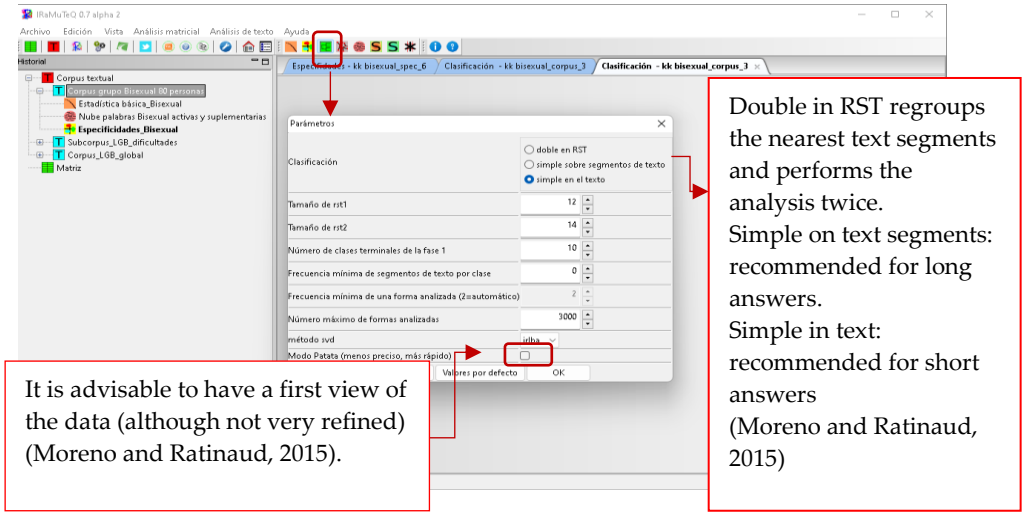

Figure 5. Selection of parameters of the Top-Down Hierarchical Ranking.

On the other hand, the CHD requires a minimum text classification of 75% (Figure 6). In other words, 75% of the texts must be organised into the classes proposed by IRaMuTeQ. If the percentage is lower than 70%, Camargo and Justo (2016; 2018) suggest disregarding the CHD and resorting to specificity analysis.

In our example (Figure 5), the classification of texts is 77.50% of the total, so that the CHD can be continued. This means that 77.50% of the texts are classified among the classes proposed by the software (subtracting 22.50% of the material outside any cluster). The resulting dendogram (Figure VI) shows how IRaMuTeQ generates these clusters, six in our case. The resulting dendogram (figure 6) shows how IRaMuTeQ generates these clusters, six in our case. In detail, it shows a first division into two subclasses: classes 5 and 2, together with class 6, on the one hand, and classes 4 and 3, and class 1, on the other hand. In a second division, classes 5 and 2, and classes 4 and 3 are separated. The division stops when IRaMuTeQ finds, in this example, six stable clusters.

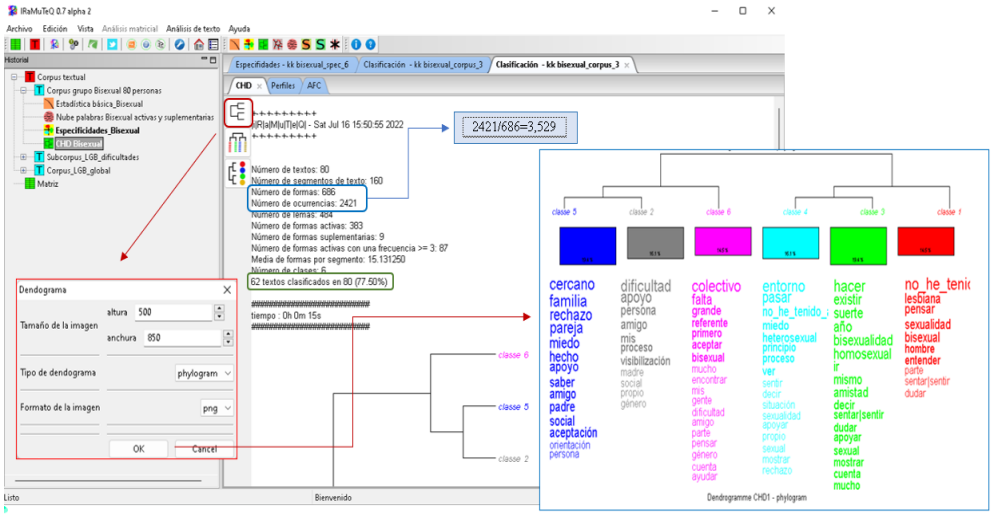

Figure 6. Preliminary CHD results and dendogram.

The "Profiles" tab (Figure 7) gives us information for each cluster (Ruiz, 2017): "n" is the order number of the form; "eff. st" is the number of text segments containing the particular word or form at least once in the corpus; "percentage" is the % of occurrence of the word in the text segments of that cluster, relative to its occurrence in the corpus); "*chi2*" represents the strength of the association of the word with its class; "type" is the type of word (noun, verb...); "Form" identifies the word in question, and "*p*" identifies the level of significance of association of the word with its class (the level of significance associated with *chi2*) (Moreno and Ratinaud, 2015).

For a descriptive analysis of the vocabulary of each class or cluster, two criteria are used simultaneously (Camargo and Justo, 2016): 1) to take into account words with a frequency higher than the mean of the set of words in the whole corpus, i.e. words with a frequency higher than no. occurrences/no. forms (see figure 6), and 2) to take into account words with $chi2 \geq 3.84$ ($p<.05$).

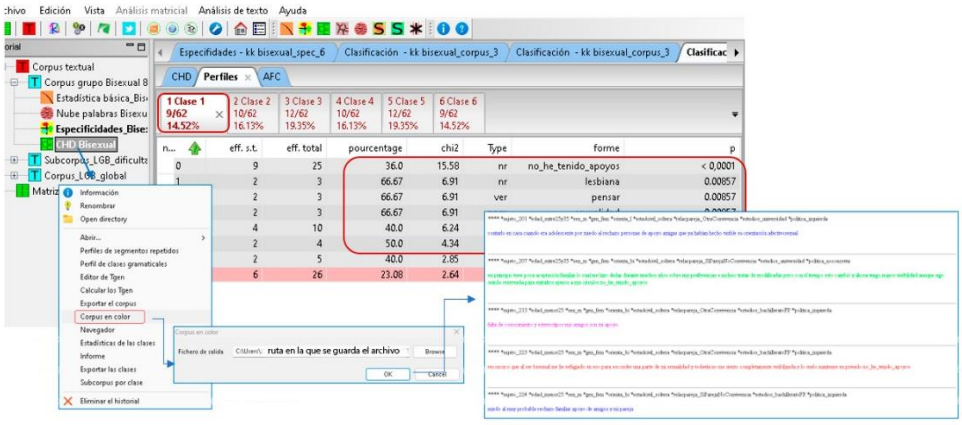

Figure 7. Details of classes, their significant forms and text fragments.

Table 1 shows the different classes, the percentage of text classified, the significant words, and their level of significance. It also shows a significant text fragment representing that class and the name attributed to each one. What name do these classes or clusters acquire? To do this, researchers will have to review in depth the most representative text fragments of each class, taking into account the statistical significance (*p*) of each form and the *chi2* value (Figure 7). These labels are added in the "Profiles" tab, by double-clicking on each class.

Table 1.

Classes, significant forms and representative texts of the classes.

Class (label)	**Significant words (*p*)**	**Representative text**
1 14,52% *"Lack of support in the process of becoming visible as a bisexual person*".	"have_not_had_supports" (*p*=.000), "lesbian" (*p*=.008), "thinking" (*p*=.008), "sexuality" (*p*=.008), "bisexual" (*p*=.012), "male" (*p*=.037)	"Being bisexual I have taken refuge in that to hide a part of my sexuality and I still don't feel fully visible and I tend to keep it private. I haven't had support" (s223sexM). "My closest environment didn't understand how I could be bisexual. I had no support" (s246sexM)
2 16,13% *"Main difficulties and support persons*".	"difficulty" (*p*=,000), "support" (*p*=,000), "person" (*p*=,003), "friend" (*p*=,003), "process" (*p*=,041)	"Accepting it was difficult as I wanted to be like everyone else. Telling my close friends and my parents was difficult. When I am with people I don't know very well I try to avoid the subject. My close friends and my parents" (s463sexM)

3 19,35% *"Differentiating bisexuality from homosexuality*".	"do" (*p*=,000), "exist" (*p*=,000), "luck" (*p*=,000), "year" (*p*=,000), "bisexuality" (*p*=,000), "homosexual" (*p*=,000), "go" (*p*=,002), "friendship" (*p*=,003), "say" (*p*=,007), "support" (*p*=,033), "sexual" (*p*=,033), "feel/sit" (*p*=,033), "show" (*p*=,033), "doubt" (*p*=,033), "account" (*p*=,033)	"During my adolescence I was very lost and alone in Chile. It was a very taboo subject, homosexuality was not talked about until very recently, and for many years only homosexual men existed. Bisexuality is completely invisible, but little by little the situation has been improving thanks to the fact that there have been people who have decided neither to hide nor to explain". (s258sexM)
4 16,13% *"The role of the environment in visibility*".	"environment" (*p*=.000), "pass" (*p*=.000), "have_not_had_supports" (*p*=.005), "fear" (*p*=.007), "heterosexual" (*p*=.017), "principle" (*p*=.041), and "process" (*p*=.041).	"I was afraid of being judged by friends and I was even more afraid that family members would stop talking to me. I felt uncomfortable coming out because when you're straight you don't have to come out" (s445sexM). "Problem of comprehension at the beginning" (s441sexM)
5 19,35% *"Fear of rejection of the closest links*".	"close" (*p*=.000), "family" (*p*=.000), "rejection" (*p*=.000), "partner" (*p*=.001), "fear" (*p*=.002), "fact" (*p*=.003), "support" (*p*=.009), "knowledge" (*p*=.016), "friend" (*p*=.026), "parent" (*p*=.033), "social" (*p*=.033), and "acceptance" (*p*=.033).	"My difficulties were probably my own, as I was slow to come to terms with it for fear of being played (...). In coming to terms with it and communicating it, I also found a certain difficulty in the fact that my mother was shocked, and the first few days she found it difficult to understand, although I have to point out that, apart from my parents, my closest family, the rest of my family does not know (...)" (s370sexM).
6 14,52% *"Role of the LGBT community*	"collective" (*p*=.000), "lack" (*p*=.000), "large" (*p*=.002), "referent" (*p*=.008), "first" (*p*=.008), "accept" (*p*=.009) and "bisexual" (*p*=.012).	"Realising I wasn't straight, coming out of the wardrobe and finding partners who didn't mind. People think that if you are bisexual you will cheat on them with someone of the opposite gender. My biggest support is my friends who are also part of the collective, almost all of them" (s225sexM).

We can use the CHD procedure as a step prior to manual categorisation (a more inductive procedure). Or we can manually code the textual material and then use the programme to complement the manual categorisation. For Santos et al. (2020), the differences between manual and automatic coding have to do with: a) the former requires more intervention on the part of the research team, and b) the more intuitive

style in coding the material than the analytical procedure of IRaMuTeQ. In any case, automatic coding does not eliminate the active participation of the research team in the understanding and significance of the clusters.

In addition, the CHD allows for a second part, which is the extraction of the factors, which are the meaningful combination of variables (figure 8). Proximity in space means correspondence between these forms. The number of factors is equal to the number of CHD classes minus 1 (Camargo and Justo, 2016). These axes show the percentage of variance they capture. In our example, the 1x2 factorial plane captures 55.22% of the variance, and it represents the variables or forms.

For the representation on the coordinate axis, Camargo and Justo (2018) suggest some options in the "AFC" menu (figure 8): in "Representation" check the option correlations; "Take the first points" by checking 100; check the option "Avoid repetitions", and "Text size proportional to chi2".

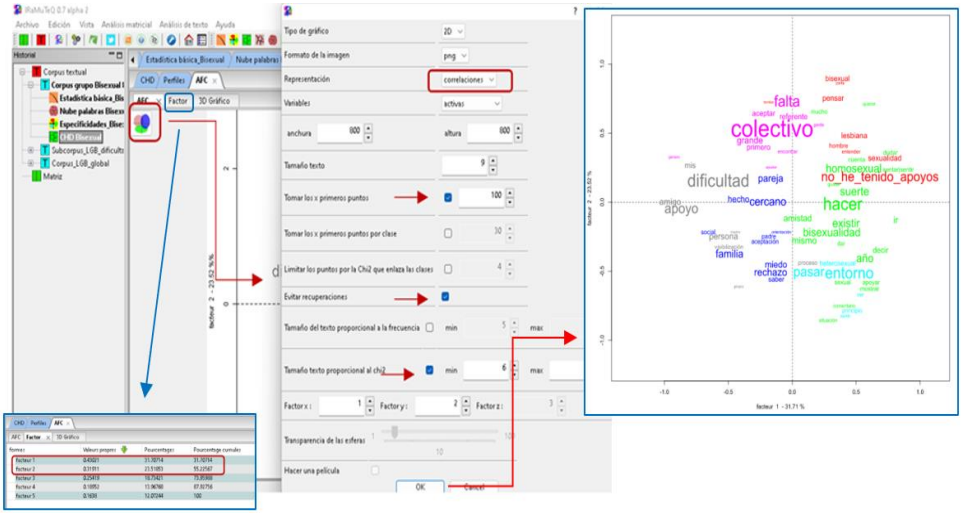

Figure 8. Extraction of factors, parameters and Cartesian representation

The result (figure 8) is a representation of the active forms on a coordinate axis (-1 and +1) in relation to their correlation (*chi2*). Class 1 has a positive relationship with factors 1 and 2; classes 2 and 5 have a negative relationship with factors 1 and 2; classes 3 and 4 have a positive relationship with factor 1 and negative with factor 2; and class 6 has a negative relationship with factor 1 and positive with factor 2. The size of the active shapes is related to their frequency, and the proximity to the correlation between them. The colours of the shapes are assigned according to the cluster they belong to. In addition, IRaMuTeQ generates by default another representation only with the labels of the classes (with the same colour as the words they agglutinate), allowing to see how the position of the classes in relation to the factors.

Advanced Steps (III): Similarity Analysis

The similarity tree (Figure 9) is based on graph theory, and graphically represents the structure of a textual corpus, differentiating the common parts and the specific parts of the coded variables (Marchand and Ratinaud, 2012).

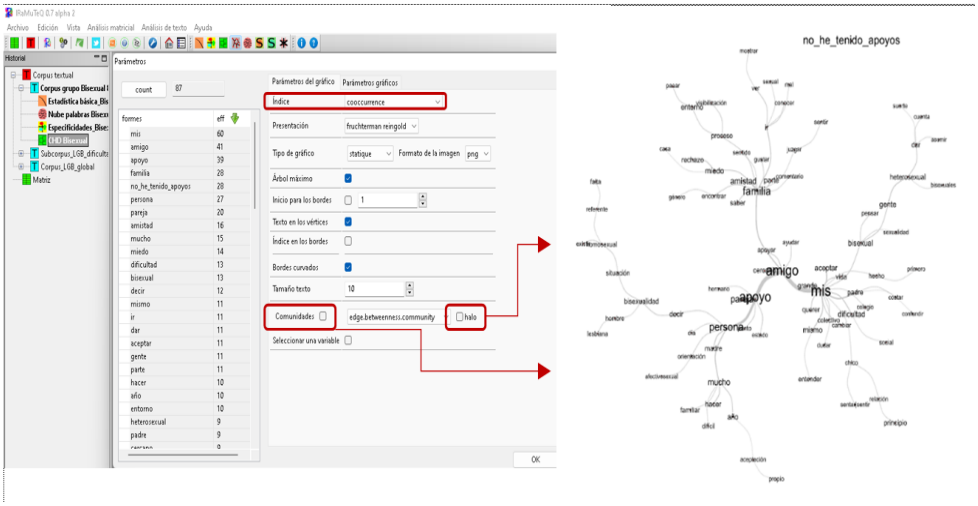

Figure 9. Similarity analysis without configuration.

The shapes/words are the vertices of the graph and the edges represent the co-occurrences between them, so that the size of the words increases with their frequency, and the links between the words are thicker the more co-occurring they are (Baril and Garnier, 2015).

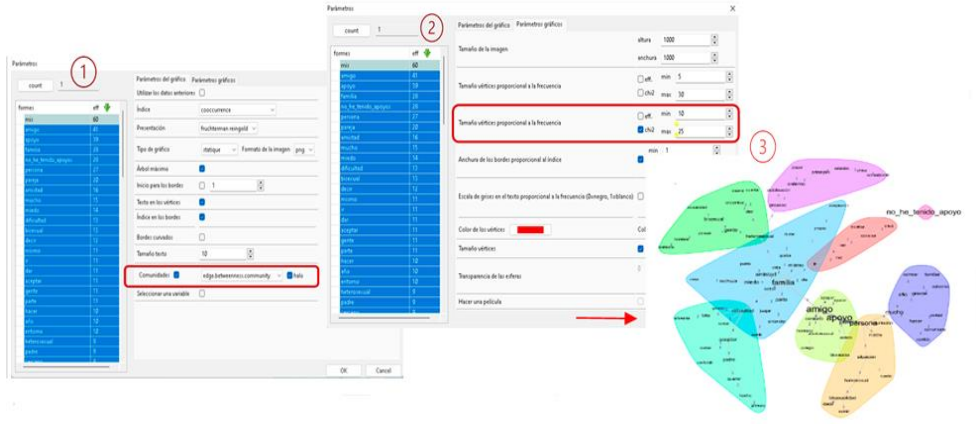

Figure 10. Maximum similarity tree with configuration.

This analysis complements the information provided by the CHD, or replaces it when the conditions for the second analysis are not met (figure 6). This first product can be refined by selecting, on the right, the words that we want to collect (by default, all are included). In addition, you can click on the "Communities" option and the "halo" option. Finally, to further improve the representation, changes are made in "Graphical settings" (figure 10).

Advanced steps (IV): sub-corpus by subject and sub-corpus by metadata

If we want to work exclusively with one theme of the data, we select the option "Subcorpus by theme", where the themes into which we have divided the textual corpus will appear, two in our example: "difficulties" and "support". If we select one, we can carry out the analysis only for that theme.

However, the more we segment the text corpus, the smaller the number of texts to be analysed, and this can be a problem. As a standard, IRaMuTeQ requires between 20 and 30 texts in the case of interviews. In the case of short answers (three or four lines) to a questionnaire, a larger number of texts is needed for the analysis to be suitable (Molina-Neira, 2017; Moreno and Ratinaud, 2015). In our example, in the subcorpus for the theme "difficulties", all analyses can be carried out. However, the subcorpus for the theme "support" only allows the analysis of specificities and similarities.

Finally, it may be of interest to analyse by metadata, i.e. by specifying a type of variable. It is possible to select more than one modality of the same variable. In our example, "*sex_h" (male sex), having a small number of participants (12 people), IRaMuTeQ only allows us to perform basic statistics and similarity analysis.

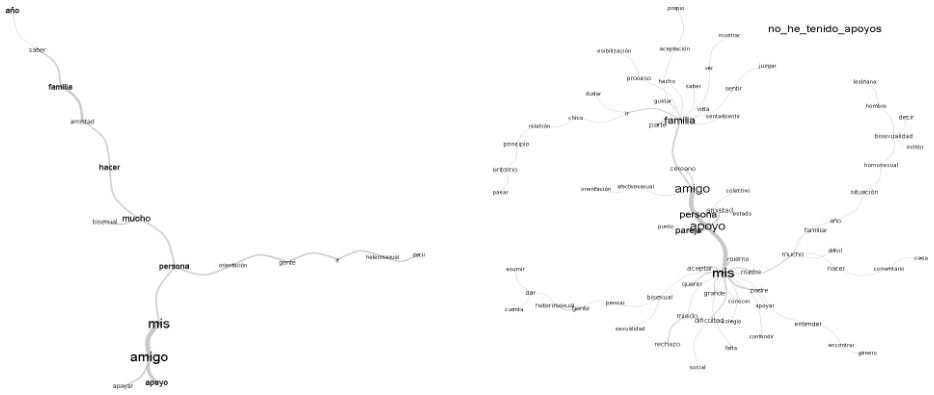

Figure 11. Difficulties and support for bisexual men (left) and women (right).

On the other hand, for "*sex_m" (female sex), the analysis of specificities and the analysis of similarities can be carried out. This procedure should make it possible to compare men's and women's representations of their difficulties and support in the process of becoming visible as bisexual (Figure XI).

Conclusions

The analysis of qualitative data can be approached from different perspectives. The lexicometric option emerges as a way of obtaining the latent meanings in a text, without the referential framework of the research team analysing the textual data. Thanks to lexicometric analysis, and on the basis of the words used and the form of expression adopted by the participants, the meaning of the discourse emerges, either of a person or a group.

Among the different existing software programmes for textual data analysis, the use of open source or free *software* is an interesting option, especially for individuals or research teams who want to use free access *software*. Among these, IRaMuTeQ stands out, with features that make it attractive for use in social and educational research, although it is necessary to take into account some particularities in its installation, which may cause initial suspicion in users who are not very experienced in the installation and use of free *software.*

The aim of this methodological article has been to present the possibilities of lexicometric analysis using the free *software* IRaMuTeQ, using the study of narratives of bisexual men and women as an example. As shown, IRaMuTeQ has reliably identified six categories or thematic classification of the textual corpus. These categories or clusters are: "*Lack of support in the process of becoming visible as a bisexual person*", "*Main difficulties and support people*", "*Differentiation of bisexuality from homosexuality*", "*Role of the environment in becoming visible*", "*Fear of rejection by the closest links*", and "*Role of the LGTB collective*".

The analysis procedure has been presented in an orderly and systematised way in order to take advantage of the potential of the software, bearing in mind that its use does not replace the participation and creativity of the researchers in obtaining results. In this sense, IRaMuTeQ organises the textual material, but it is up to the research team to prepare the textual material thoroughly, as well as to decide whether to carry out the different analyses available by previous thematic areas, or with the textual material as a whole. Moreover, the grouping of the narrative material (the clusters) is given meaning by the research team's exercise of naming them, thus giving meaning to the manifest content (the words) involved in that cluster. Although the programme performs an automatic analysis, it does not replace the researcher's creativity, since it is up to them to give meaning to the material grouped by IRaMuTeQ.

References

Ander-Egg, E. (2011). *Learning to do research. Basic notions for social research*. Editorial Brujas.

Ary, D., Jacobs, L. C., & Sorensen, C. (2010). *Introduction to Research in Education*. Wadsworth Publishing.

Ávalos, M. A., Merma, G., Hernández-Amorós, M. J., Urrea-Solano, M. E., and Aparicio, M. P. (2018). Redesign of the Tutorial Action Plan based on the focus group. The case of the Faculty of Education. In: R. Roig-Vila (Ed.), *El compromiso académico y social a través de la investigación e innovación educativas en la Enseñanza Superior (*pp. 911-921). Octaedro.

Baril, E., and Garnier, B. (2015). *IRaMuTeQ 0.7 alpha 2.* http://www.iramuteq.org/documentation/fichiers/Pas%20a%20Pas%20IRAMUTEQ_0.7alpha2.pdf

Baccalá, N., and de la Cruz, M. (2000). The importance of textual statistics applied to the study of teaching conceptions. In: M. Rajman. and J. C. Chappelier (eds.) 5es Journées Internationales d'Analyse Statistique des Données Textuelles (pp. 519-524). Switzerland. https://lexicometrica.univ-paris3.fr/jadt/jadt2000/pdf/39/39.pdf

Becue, M., Lebart, L., and Rajadell, N. (1992). Statistical analysis of textual data. Reading according to primary school children. *Anuario de Psicología*, 55, 7-22. https://www.raco.cat/index.php/AnuarioPsicologia/article/view/61168

Braun, V., and Clarke, V. (2006). *Using thematic analysis in Psychology. Qualitative Research in Psychology, 3(*2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Camargo, B. V., and Justo, A. M. (2013). IRAMUTEQ: Um Software Gratuito para Análise de Dados Textuais. *Temas em Psicologia, 21*(2), 513-518. http://www.redalyc.org/articulo.oa?id=513751532016

Camargo, B. V., and Justo, A. M. (2016). *Tutorial para uso do sotfware IRaMuTeQ*. UFSC Brasil.

Carvajal, D. (2002). The Craftswoman's Tools. Critical Aspects in the Teaching and Learning of CAQDAS. *Forum: Qualitative Social Research, 3*(2). https://doi.org/10.17169/fqs-3.2.853

Chandra, Y., and Shang, L. (2019). *Qualitative Research Using R: A Systematic Approach*. Springer.

Chernobilsky, L. B. (2009). The use of the computer as an auxiliary in the analysis of qualitative data. In: I. Vasilachis (ed.), *Estrategias de investigación cualitativa* (pp. 239-273). Gedisa.

Díaz Herrera C. (2018). Qualitative research and thematic content analysis. Intellectual orientation of Universum journal. *General Journal of Information and Documentation*, *28*(1), 119-142. https://doi.org/10.5209/RGID.60813

Farías, L., and Montero, M. (2005). On transcription and other craft aspects of qualitative research. *International Journal of Qualitative Methods, 4*(1). https://sites.ualberta.ca/~iiqm/backissues/4_1/pdf/fariasmontero.pdf

García Montejo, S. (2015). Methodological aspects of qualitative research. In L. Abero, L. Berardi, A. Capocasales, S. Garcia Montejo and R. Rojas, *Educational research. Opening doors to knowledge (*pp. 101-118). CLACSO.

Gibbs, G. R. (2012). *The analysis of qualitative data in qualitative research*. Morata.

Gil, J., García, E., Rodríguez, G., and Corrales , A. (1994) Análisis estadístico de datos cualitativos textuales: enfoque lexicométrico. *Educational Research Journal*. 23, 510-514. http://hdl.handle.net/11162/185311

Glaser, B. G., and Strauss, A. L. (1967). *The Discovery of Grounded Theory. Strategies for Qualitative Research*. Aldine.

Gonçalves, J., Sales, J. P., Moreno, M. M., & Rolim, M. L. (2021). The Impacts of SARS-CoV-2 Pandemic on Suicide: A Lexical Analysis. *Frontiers in Psychiatry*, 12, 593918. https://doi.org/10.3389/fpsyt.2021.593918

González-Teruel, A. (2015). Methodological strategies for user research in social media: content analysis, grounded theory and discourse analysis. *Profesional de la Información*, *24*(3), 321-328. https://doi.org/10.3145/epi.2015.may.12

Gourlay, S. (2019). *Preparing IRaMuTeQ input files*. Kingston University.

Hart, T., and Achterman, P. (2017). Qualitative Analysis Software (ATLAS.ti/Ethnograph/MAXQDA/NVivo). In: C. S. Davis and R. F. Potter (Eds.), *The International Encyclopedia of Communication Research Methods*. John Wiley and Sons.

Hernández-Sampieri, R., and Mendoza, C. P. (2018). *Research methodology. Quantitative, qualitative and mixed routes*. McGraw-Hill.

Huber, G., and Marcelo, C. (1990): More than word retrieval and frequency counting: Computer aided analysis of qualitative data. *Teaching*, 8, 69-84.

Jordà, G., Pàmies, C., and Vicens, C. (2000) El análisis lexicométrico en literatura. Methodological proposal and applications. In M. L. Casal, G. Conde, J. Lago, L. Pino and N. Rodriguez (eds.), *La lingüística francesa en España camino del siglo XXI* (pp. 610-617). Arrecife.

Kalpokaite, N., and Radivojevi, I. (2019). Demystifying Qualitative Data Analysis for Novice Qualitative Researchers. *The Qualitative Report, 24*(13), 44-57. https://nsuworks.nova.edu/tqr/vol24/iss13/5

Loubère, L., and Ratinaud, P. (2014). *Documentation IRaMuTeQ 0.6 alpha 3 version 0.1.* http://www.iramuteq.org/documentation/fichiers/documentation_19_02_2014.pdf

Marchand, P. and Ratinaud, P. (2012). L'analyse de similitude appliquée aux corpus textuels: les primaries socialistes pour l'élection présidentielle française (septembre-octobre 2011). Proceedings of the *XI Journées Internationales d'Analyse statistique des Données Textuelles* (pp. 687-699). Liège. Belgium.

Marina, J., & Feliz, T. (2018). Perceptions of information seeking and health education in virtual environments in Spanish. *Revista Española de Salud Pública*, *92*, https://tinyurl.com/26jqwe9s

Massot, I., Dorio, I., and Sabariego, M. (2004). Strategies for collecting and analysing information. In: R. Bisquerra (coord.), *Metodología de la investigación educativa* (pp. 329-368). La Muralla.

Miles, M. B., and Huberman, A. M. (1994). *Qualitative Data Analysis. An Expanded Sourcebook*. Sage Publications.

Molina-Neira, J. (2017). *Tutorial for text analysis with IRaMuTeQ software*. University of Barcelona.

Morales, C. (2021). Use of IRaMuTeQ software. In: M. A. Casillas, J. Dorantes and C. Ortiz-Blanco (coord.), *Representaciones sociales, educación y análisis cualitativo con IRaMuTeQ* (pp. 23-38). Universidad Veracruzana.

Moreno, M., and Ratinaud, P. (2015). *IRaMuTeQ user manual.* http://www.iramuteq.org/documentation/fichiers/guia-iramuteq

Peralta, N., Castellaro, M., and Santibáñez, C. (2019) Analysing Textual Data As a Methodology to Approach Argumentation: A Research Work Among Undergraduate Students from Chilean Universities**.** *Íkala, Revista de Lenguaje y Cultura, 25*(1), 209-227**.** https://doi.org/10.17533/udea.ikala.v25n01a02

Ratinaud, P. (2018). Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ. In D. F. Iezzi, L. Celardo, & M. Misuraca (Eds.), *JADT' 2018, Proceedings of the 14th international conference on statistical analysis of textual data* (pp. 616-625). Universitalia. http://lexicometrica.univ-paris3.fr/jadt/JADT2018/actes-jadt18.pdf

Richards, L., and Richards, T. (2002). From filing cabinet to computer. In: A. Bryman and R. G. Burgess, (Eds.), *Analyzing qualitative data* (pp. 146-172). Routledge.

Romero-Pérez, I., Alarcón-Vásquez, Y., and García-Jiménez, R. (2018). Lexicometrics: an applied approach to the redefinition of concepts and identification of thematic units. *Biblios*, 71, 68-80. https://doi.org/10.5195/biblios.2018.466. https://doi.org/10.5195/biblios.2018.466

Ruiz, A. (2017). *Working with IRaMuTeQ: Guidelines*. Universitat de Barcelona. http://hdl.handle.net/2445/113063

Ryan, G. W., and Bernard, H. R. (2000). Data management and analysis methods. In: N. K. Denzin and Y. S. Lincoln (Eds.), Handbook of Qualitative Research (pp. 769-802). S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 769-802). Sage Publications.

Sabariego, M. (2004). The research process (part 2). In: B. Rafael (coord.), *Metodología de la investigación educativa (*pp. 127-163). La Muralla.

Saldaña, J. (2013). *The Coding Manual for Qualitative Researchers*. SAGE.

Sandín, M. P. (2003). *Qualitative Research in Education. Fundamentals and traditions*. McGraw-Hill.

Santos, I. C., Rosário, V. M., Amaral-Rosa, M. P., Moreira, L., and Ramos, M. G. (2020). Handcrafted and Software-Assisted Procedures for Discursive Textual Analysis: Analytical Convergences or Divergences? In: A. P. Costa, L. P Reis, and A. Moreira (Eds.), *Computer Supported Qualitative Research (*pp. 189-205). Springer.

Sbalchiero, S. (2018). Topic Detection: A Statistical Model and a Quali-Quantitative Method. In: A. Tuzzi (ed.), *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences* (pp. 189-209). Springer.

Souza, M. A. R., Wall, M. L., Thuler, A. C., Lowen, I. M. V., & Peres, A. M. (2018). The use of IRAMUTEQ software for data analysis in qualitative research. *Revista Da Escola de Enfermagem Da USP*; 52, e03353. http://dx.doi.org/10.1590/S1980-220X2017015003353

Taylor, S. J., and Bogdan , T. (1987). *Introduction to qualitative research methods*. Paidós.

Trigueros-Cervantes, C., Rivera-García, E., & Rivera-Trigueros, I. (2018). *Conversational and narrative techniques. Qualitative research with NVivo software*. University of Granada. Andalusian School of Public Health. https://www.easp.es/wp-content/uploads/publicaciones/UGR-EASP_Libro_Cualitativa-NVivo_12.pdf

van Dijk, T. A. (1999). Critical discourse analysis. *ANTHROPOS*, 186, 23-36.

Date of receipt: 10 March 2023.
Review date: 28 May 2023.
Date of acceptance: 9 October 2023.